casa Engineering Research Center for
Collaborative Adaptive Sensing of the Atmosphere

On Characteristics and Modeling of P2P Resources with Correlated Static and Dynamic Attributes

H. M. N. Dilum Bandara and Anura P. Jayasumana

Electrical and Computer Engineering,

Colorado State University, Fort Collins, CO.

dilumb@engr.colostate.edu

University of Massachusetts Amherst

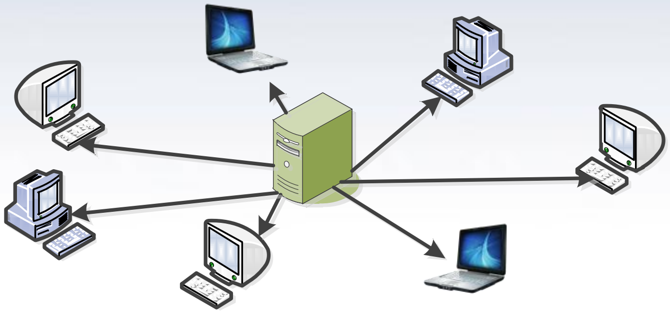
University of Oklahoma

Colorado State University

University of Puerto Rico Mayaguez

CASA is primarily supported by the Engineering Research Centers Program of the National Science Foundation under NSF award number 0313747.

Motivation

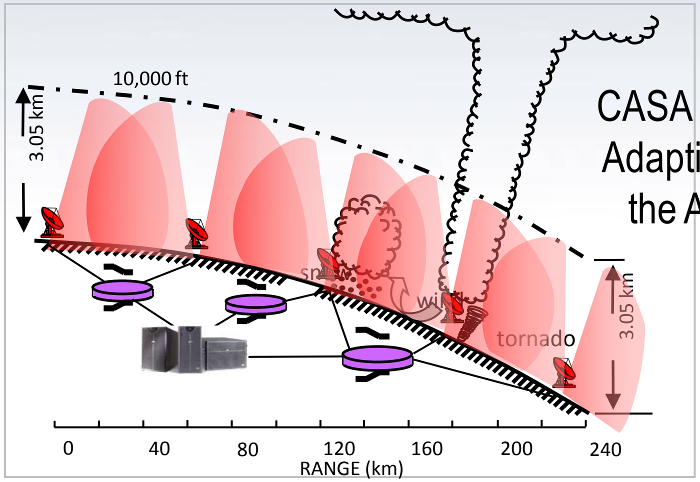


Desktop grids

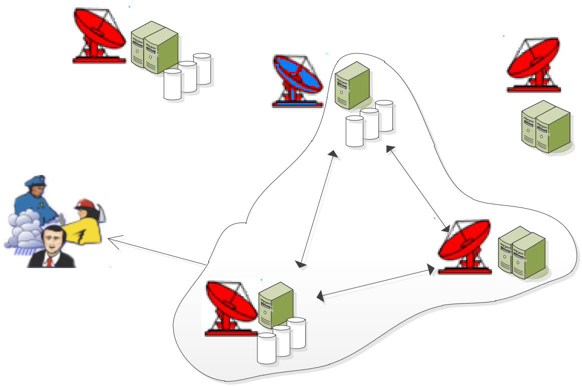
- Throughput
- Static attributes
  - CPU speed, architecture, GPUs

Community (P2P) clouds

- QoE & QoS
- Dynamic attributes
  - Free CPU, bandwidth



CASA (Collaborative Adaptive Sensing of the Atmosphere)



Radar networks

- QoS & latency
- Static & dynamic attributes
  - CPU speed, free CPU, bandwidth

Contributions

Mechanism to generate realistic, synthetic traces of P2P resources with multivariate static & dynamic attributes

- Characteristics & models of resources are essential in design, validation, & performance analysis
- Neither practical nor economical to capture large-scale & high-resolution datasets

- Enable large-scale performance studies of resource discovery solutions, job schedulers, applications, etc.

Objectives

- Understand characteristics & model multi-attribute resources
  - Static attributes [Heien, 2011]
  - Static & dynamic attributes, & queries [Bandara, 2011]
  - Existing performance studies assume
    - i.i.d attributes, uniform/Zipf's distribution of attributes, ignored dynamic attributes, replication of small datasets, etc.
    - Not valid
- Generate large synthetic datasets that preserve statistical properties of real-life systems
  - Large number of resources & attributes
  - Preserve correlation, temporal patterns
  - Valid from few minutes to few days
  - Dataset neutral

- E. Heien et al., "Correlated resource models of Internet end hosts," ICDCS '11, June 2011.
- H.M.N.D. Bandara & A.P. Jayasumana, "Characteristics of multi-attribute resources/queries and implications on P2P resource discovery," AICCSA '11, Dec. 2011.

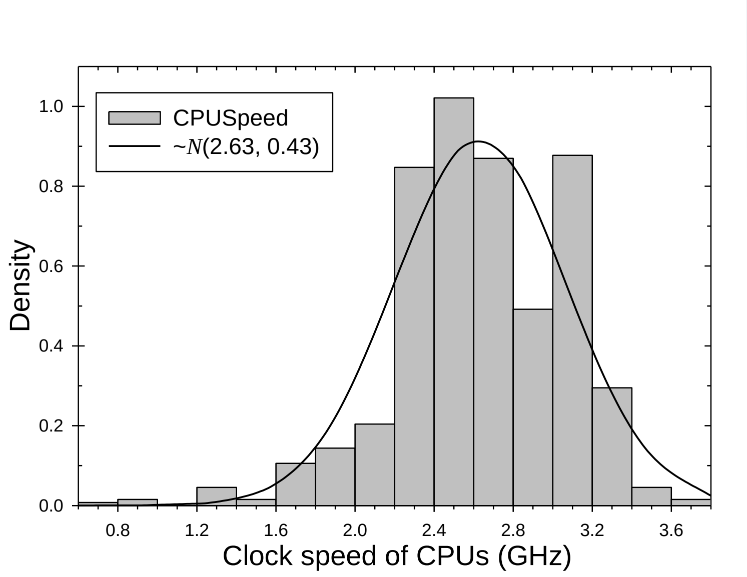
Dataset

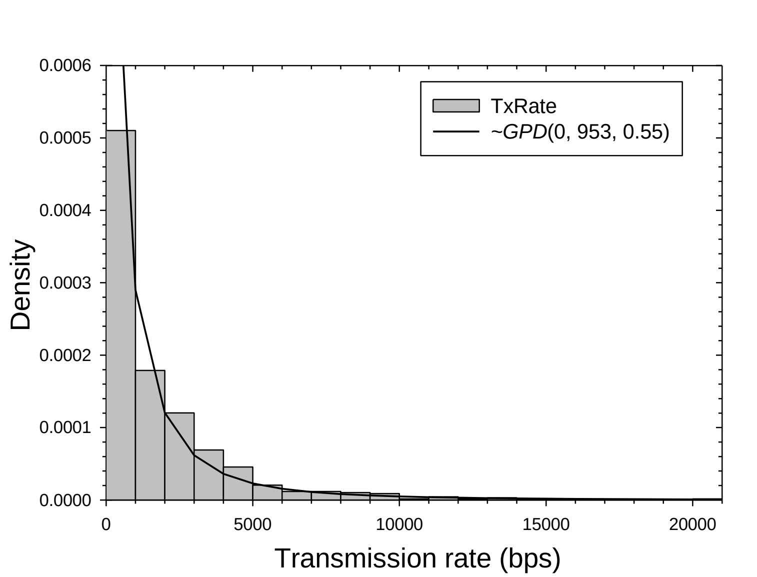
- PlanetLab node data
  - Global research network for developing new network services, protocols, & applications
  - Reflects many characteristics of Internet-based distributed systems
    - Heterogeneity, multiple end users, dynamic nodes, & global presence
    - Used to evaluate many preliminary P2P protocols & applications
- Rich dataset with
  - 12 static & 34 dynamic attributes
  - 5 min samples
  - 500-700 active nodes
  - Collected between Nov 1 to 15, 2010
- Currently collecting public & campus datasets

Resource Model

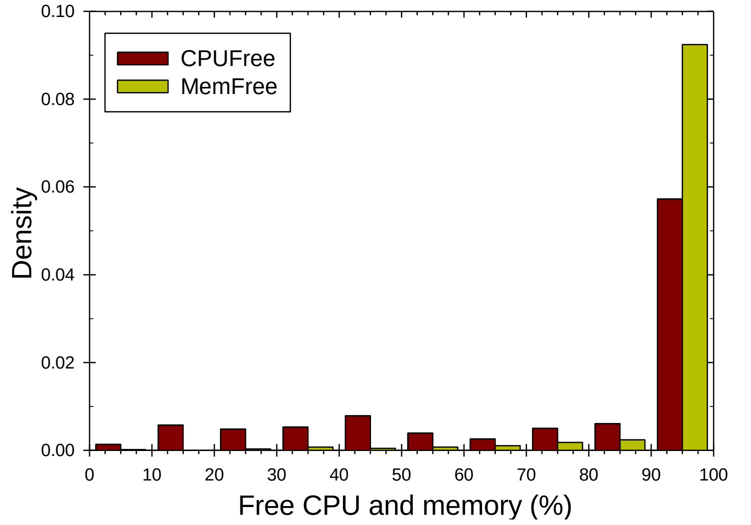
Attribute	Description
CPUSpeed	CPU clock speed in GHz. Provides in-sight on relative computing power of a node
NumCores	No of processor cores. How much parallelism in processing is possible?
MemSize	Size of volatile memory in GB
CPUFree	(100 – CPU utilization)%. To what extent the CPU(s) is available for processing. Average is given for multiple cores
MemFree	Free user-level memory as a %. Amount of memory is available for user processes
DiskFree	Free disk space in GB.
1MinLoad	1 min exponentially weighted moving average of number of active processes competing or waiting for CPU. How long a user process has to wait?
TxRate	Average transmission rate in bps. In conjunction with bandwidth limit specified by most nodes, it provides insight on amount of available bandwidth
RxRate	Average receive rate in bps

Resource Characteristics – Distributions

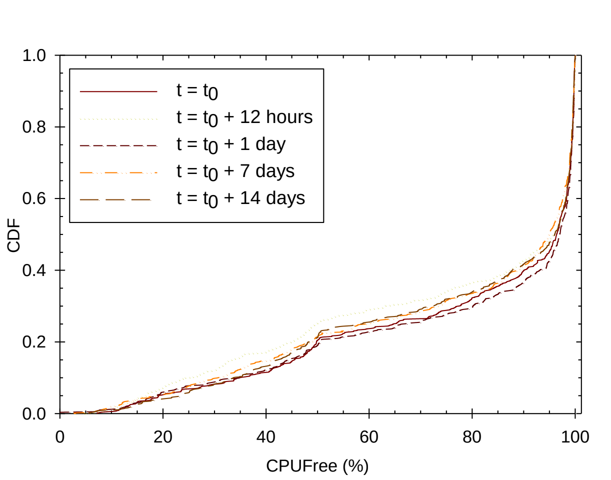


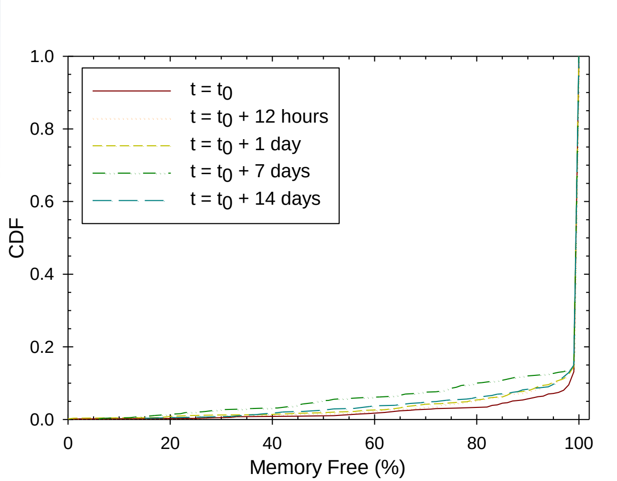


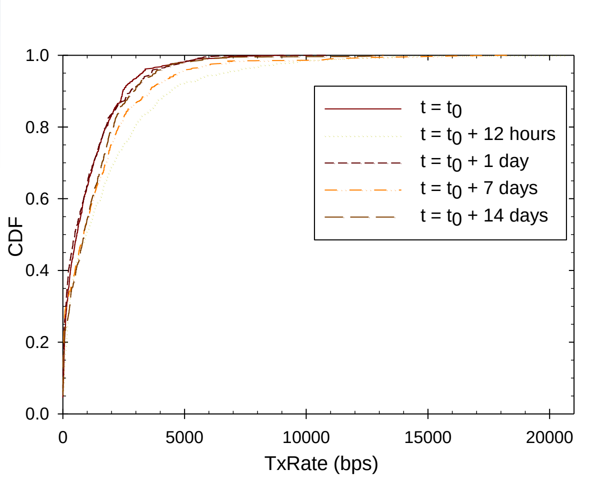
- Resources satisfy a mixture of probability distributions
  - Gaussian – CPUSpeed, MemSize, DiskFree
  - Pareto – TxRate, RxRate
- Highly skewed distributions
  - CPUFree & MemFree



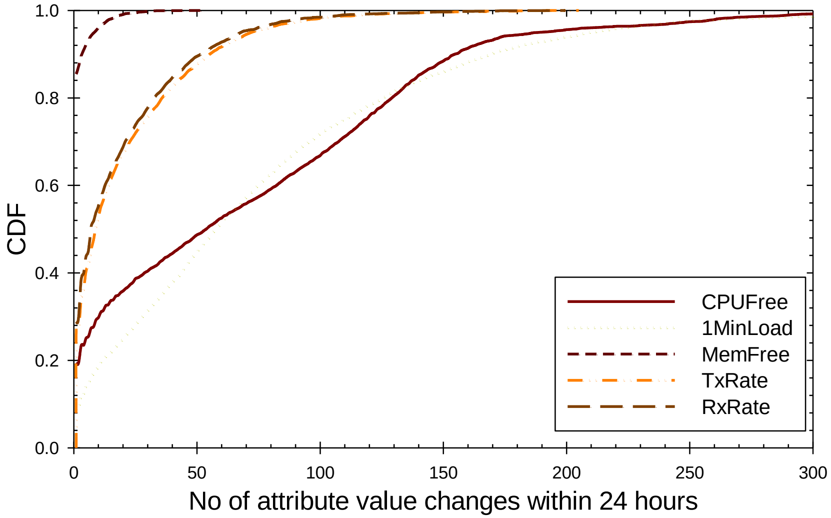
Dynamic Attributes at Different Times







- Distributions of dynamic attributes are stable over days
- Dynamic attributes & their rate of change fits Pareto distribution
  - Some attributes/nodes change frequently
  - Many status updates



Thresholds: *CPUFree* = *MemFree* = ± 10%, *1MinLoad* = ± 2, *TxRate* = *RxRate* = ± 1 Kbps

Resource Characteristics – Correlation

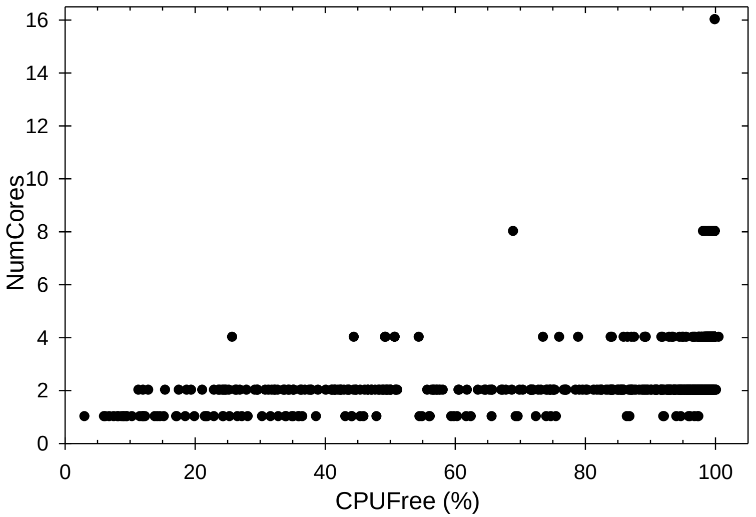
Pearson’s correlation coefficient

	CPUSpeed	NumCores	CPUFree	1MinLoad	MemSize	MemFree	DiskFree	TxRate
NumCores	-0.09							
CPUFree	0.02	0.48						
1MinLoad	0.03	-0.31	-0.57					
MemSize	0.06	0.28	0.26	-0.25				
MemFree	0.13	0.21	0.31	-0.35	0.25			
DiskFree	-0.09	0.46	0.37	-0.29	0.54	0.23		
TxRate	0.08	-0.23	-0.26	0.24	-0.12	-0.17	-0.12	
RxRate	0.10	-0.23	-0.30	0.35	-0.13	-0.20	-0.16	0.85

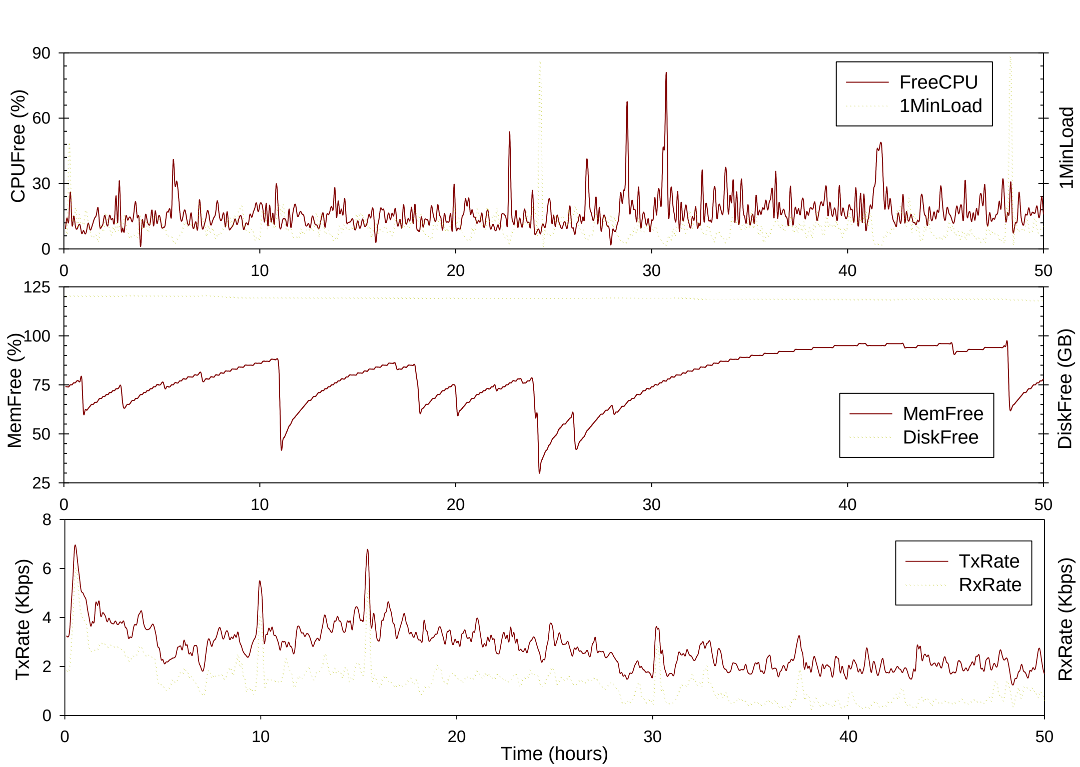
Spearman’s ranked correlation coefficient ρ

	CPUSpeed	NumCores	CPUFree	1MinLoad	MemSize	MemFree	DiskFree	TxRate
NumCores	0.04							
CPUFree	-0.07	0.67						
1MinLoad	0.10	-0.42	-0.72					
MemSize	0.03	0.37	0.37	-0.33				
MemFree	-0.07	0.37	0.37	-0.38	0.53			
DiskFree	-0.20	0.60	0.52	-0.41	0.44	0.44		
TxRate	0.06	-0.35	-0.39	0.30	-0.07	-0.20	-0.29	
RxRate	0.07	-0.33	-0.42	0.41	-0.11	-0.21	-0.29	0.86

- Complex correlation among attributes
- Correlation between
  - Static-dynamic attributes
  - Dynamic-dynamic attributes

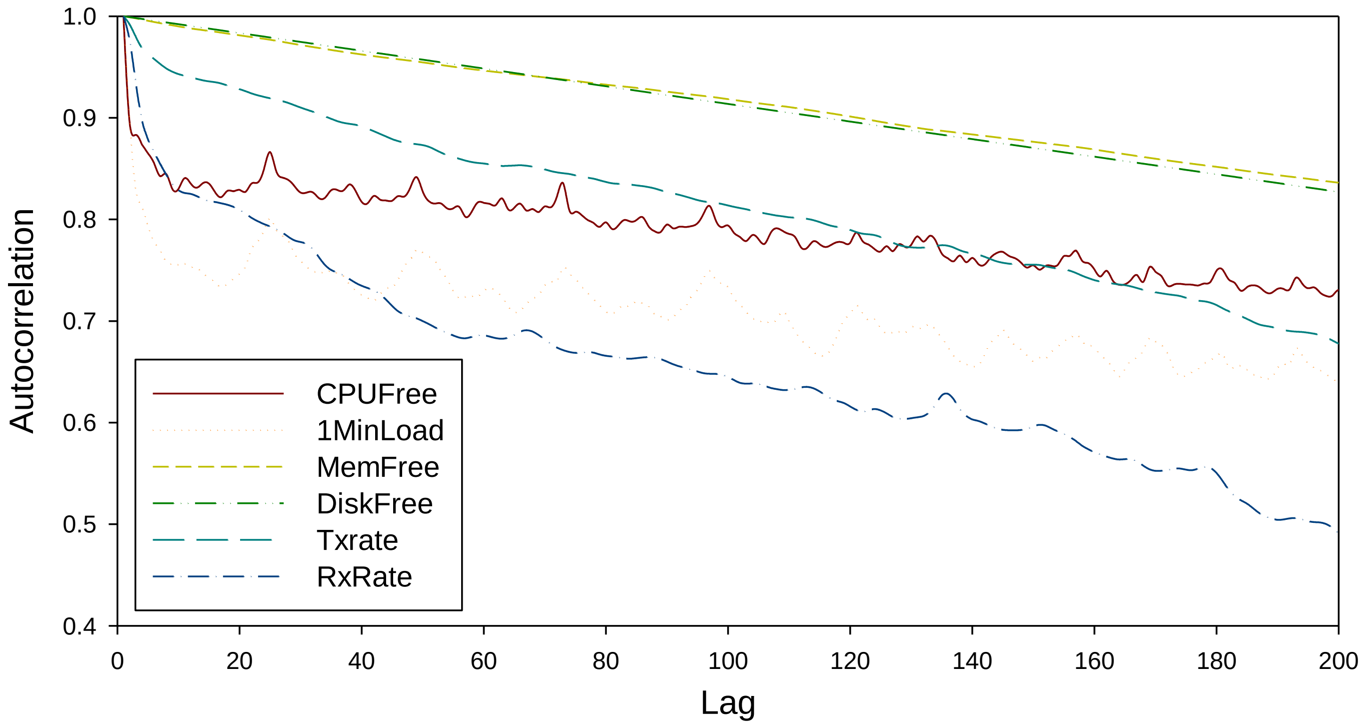


Dynamic Attributes – Contemporaneous Correlation



- Contemporaneous correlation among time series of dynamic attributes
- Specific temporal pattern in MemFree
- Temporal patterns need to be preserved

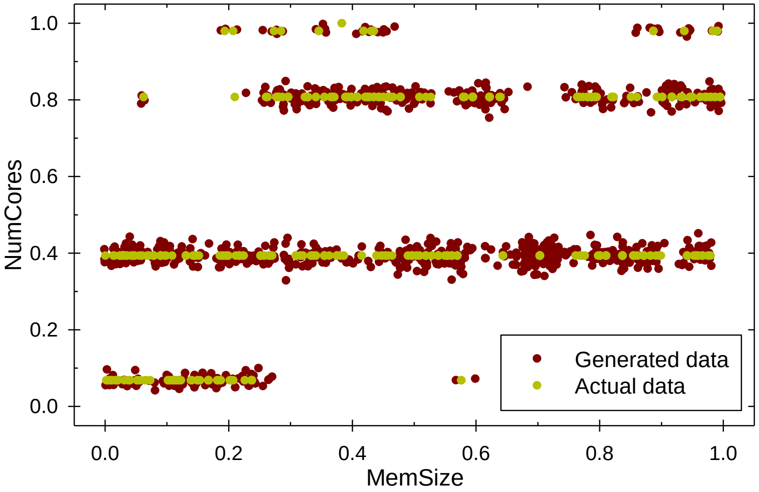
Dynamic Attributes – Autocorrelation



- High autocorrelation in DiskFree & MemFree
- No noticeable change in DiskFree
- Temporal patterns need to be preserved

Modeling Static Attributes

- Need to preserve correlation
  - Attribute values can’t be randomly drawn from marginal distributions
  - Pearson’s correlation matrix is insufficient
- Copulas capture complex correlations
  - Functions that couple multivariate distributions to their marginals
  - Multivariate joint distribution defined on d-dimensional unit cube s.t. marginal distribution u_i is $\sim uniform(0, 1)$
  - $F(u) = C(F_1(u_1), \ldots, F_d(u_d))$
- Empirical copulas support complex/unknown distributions & correlations
  - $C_n\left(\frac{i}{n}, \frac{j}{n}\right) = \frac{\text{No of pairs}(x, y)\text{ s.t. } x \le x_{(i)} \text{ and } y \le y_{(j)}}{n}$
  - $x_{(i)}$ ordered statistics of x
  - No need to find distribution of attributes



Modeling Dynamic Attributes

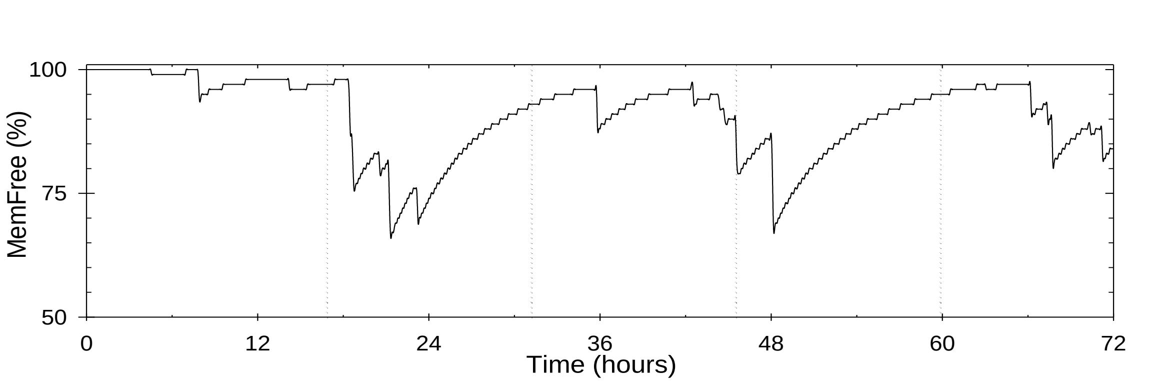
- Specific temporal patterns in time series → can't draw values randomly
- Contemporaneous correlation → can't draw independently
- Goal – Not to predict future behavior, but to generate nodes with similar overall characteristics
  - Not necessary to fit a model
- Build a library of time series segments
  - Pick the most distinct pattern & split according to structural changes
    - Preserve distinct temporal patterns
  - Split other time series at same position & replay segments together

Modeling Dynamic Attributes (cont.)

$$y_i = x_i^T \beta_i + u_i \quad i = 1, \ldots, n$$

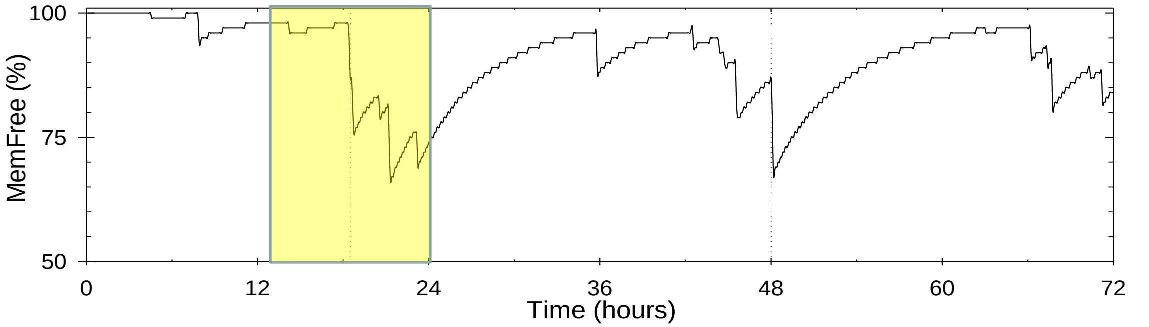
Check for Null Hypothesis that

H_0: $\beta_i = \beta_0$, $i = 1, \ldots, n$



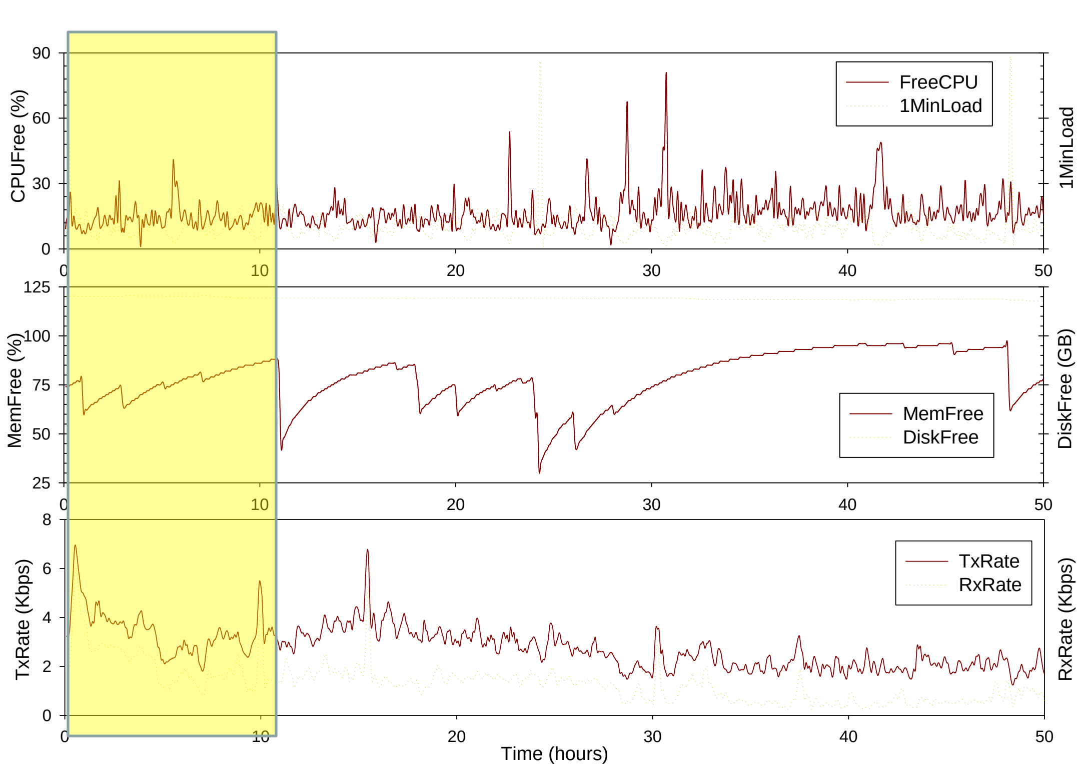
Sliding Window

w = 20, Δ = 30%



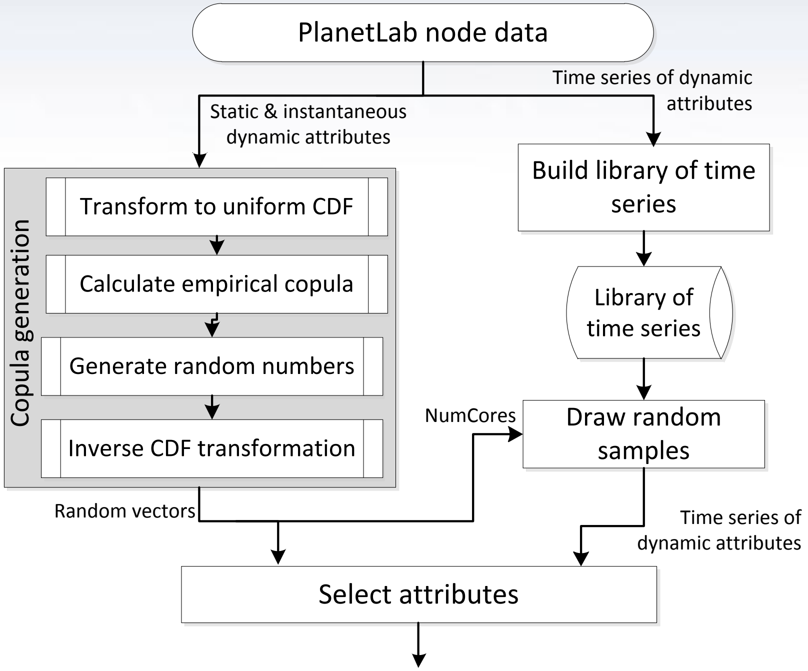
- Initial approach – R *strucchange* package
- Better approach – Sliding window (w) looking for significant change in average value (Δ) of 2 halves of the window

Dynamic Attributes – Contemporaneous Correlation



- Split other time series at same position & replay segments together
- Concatenate segments to form longer sequences
- Segments are index by static attributes

Resource Generation Tool



PlanetLab node data

Static & instantaneous dynamic attributes

Time series of dynamic attributes

Copula generation

Transform to uniform CDF

Calculate empirical copula

Generate random numbers

Inverse CDF transformation

Build library of time series

Library of time series

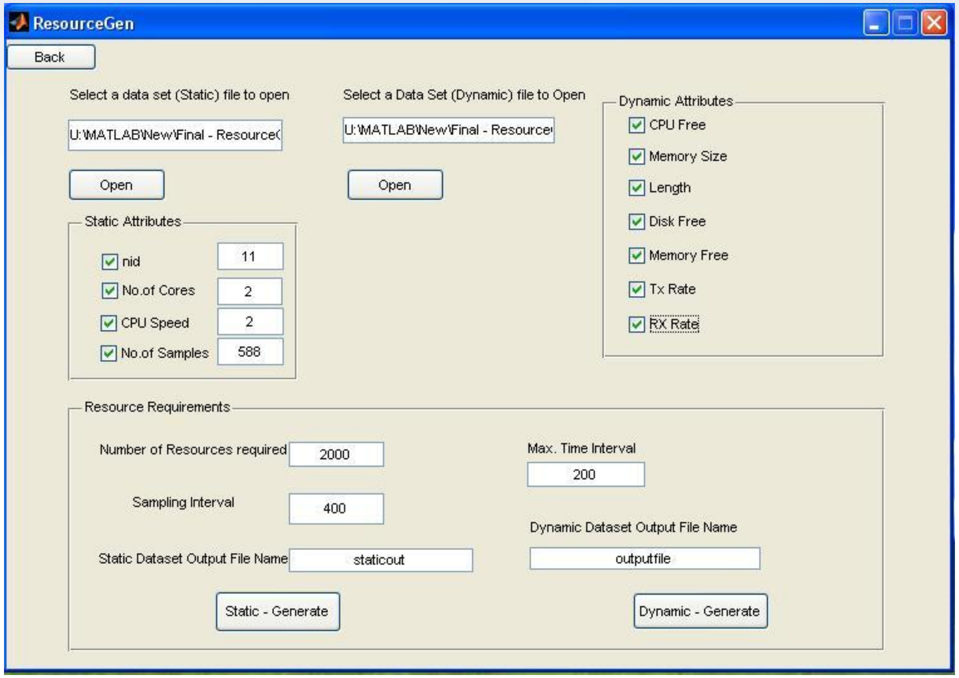
NumCores

Draw random samples

Random vectors

Time series of dynamic attributes

Select attributes



Select a data set (Static) file to open: U:\MATLAB\New\Final - Resource(

Select a Data Set (Dynamic) file to Open: U:\MATLAB\New\Final - Resource

Static Attributes

- ☑ nid: 11
- ☑ No.of Cores: 2
- ☑ CPU Speed: 2
- ☑ No.of Samples: 588

Dynamic Attributes

- ☑ CPU Free
- ☑ Memory Size
- ☑ Length
- ☑ Disk Free
- ☑ Memory Free
- ☑ Tx Rate
- ☑ RX Rate

Resource Requirements

Number of Resources required: 2000

Sampling Interval: 400

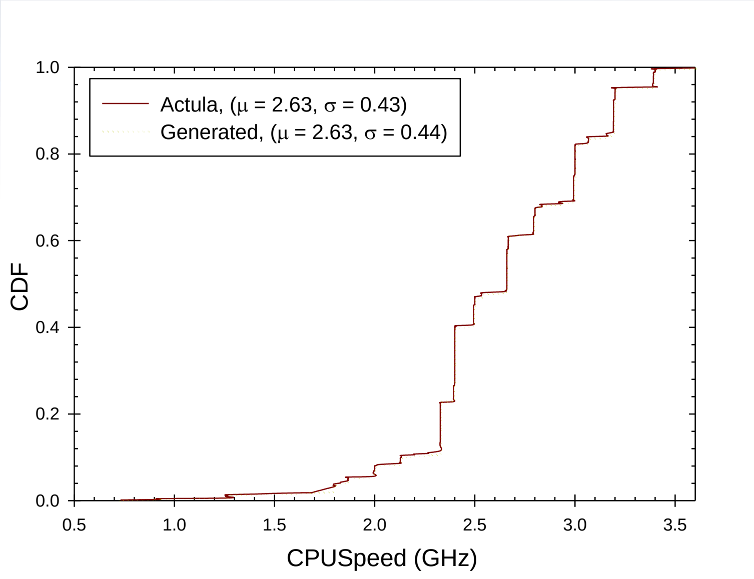
Max. Time Interval: 200

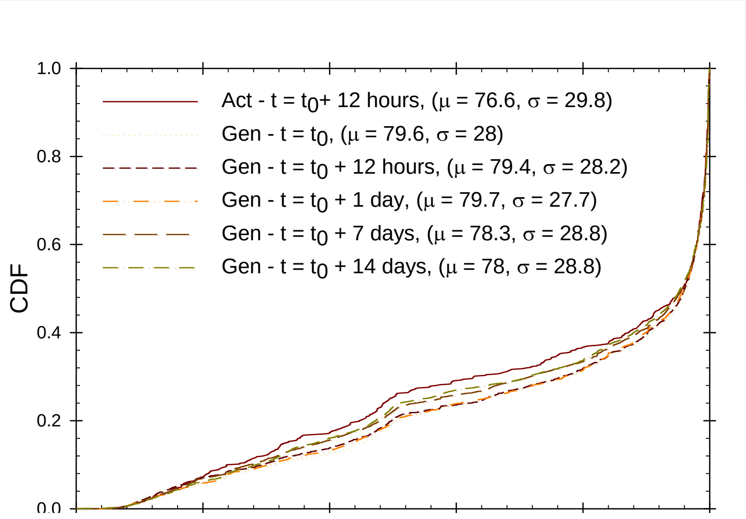
Static Dataset Output File Name: staticout

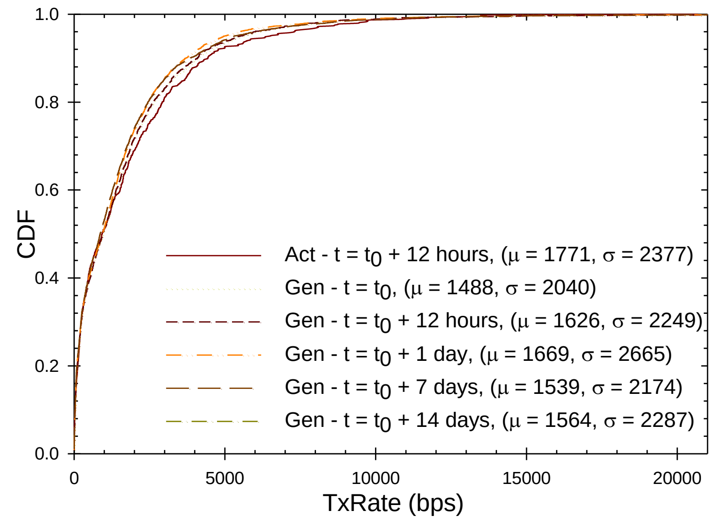
Dynamic Dataset Output File Name: outputfile

- NumCores establish correlation between static & dynamic
- Generate synthetic traces with n nodes, a_s static & a_d dynamic attributes over a given time t
- Available – www.cnrl.colostate.edu/Projects/CP2P/

Resource Generation – Validation

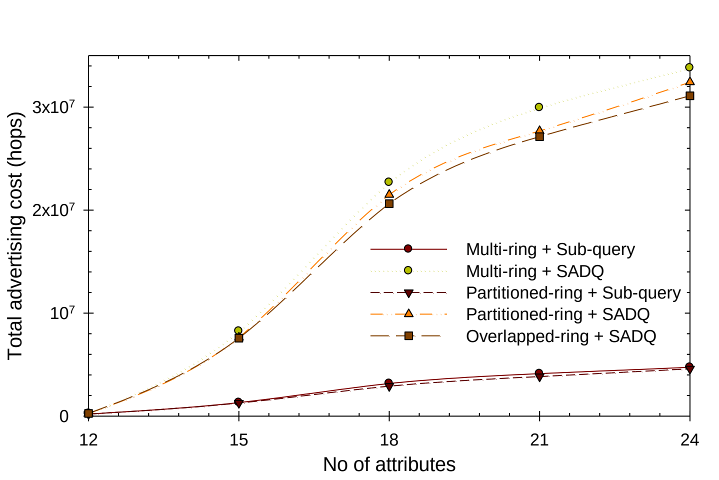


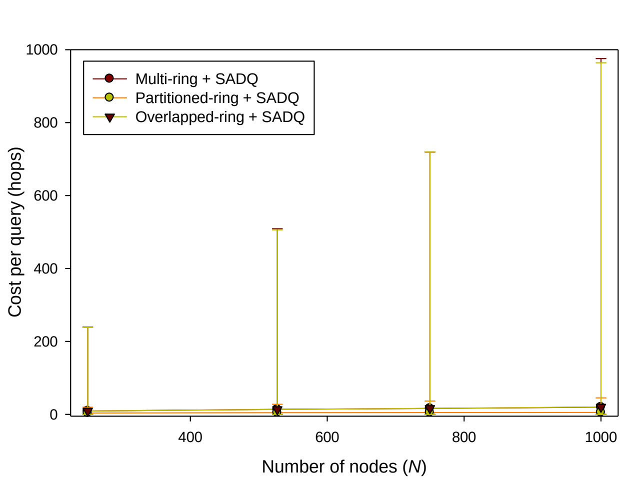




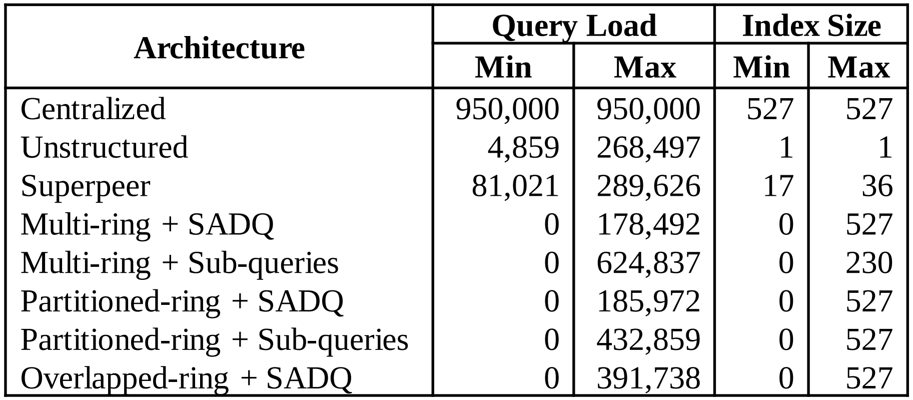
- Using 300 nodes over a week → generated 5,000 nodes over 2 weeks
- Satisfy Kolmogorov-Smirnov (KS) test with a significance level of 0.05
- Statistically accurate data

Application of Tool – Performance Analysis of P2P Resource Discovery Solutions





- Used to study existing P2P resource discovery solutions
- Identify issues such as
  - High cost of updating/advertising dynamic attributes
  - High resource discovery cost $O(N)$
  - Load balancing



Architecture	Query Load		Index Size	
	Min	Max	Min	Max
Centralized	950,000	950,000	527	527
Unstructured	4,859	268,497	1	1
Superpeer	81,021	289,626	17	36
Multi-ring + SADQ	0	178,492	0	527
Multi-ring + Sub-queries	0	624,837	0	230
Partitioned-ring + SADQ	0	185,972	0	527
Partitioned-ring + Sub-queries	0	432,859	0	527
Overlapped-ring + SADQ	0	391,738	0	527

H.M.N.D Bandara and A.P. Jayasumana, "Evaluation of P2P Resource Discovery Architectures Using Real-Life Multi-Attribute Resource and Query Characteristics," IEEE CCNC '12, Jan. 2012.

Conclusions

- Technique to generate vectors of static attributes & multivariate time series of dynamic attributes
  - Supports complex/mixed distribution of attributes
  - Works with other multivariate datasets
  - Can be applied to collaborative P2P, cloud computing, community clouds, desktop grids, etc.
  - Evaluate scalability of applications, resource discovery solutions, & job schedulers far beyond that's possible with existing test beds
- Future work
  - Support new datasets being collected
  - Support node failures/availability
  - Model multi-attribute queries
  - Build efficient multi-attribute resource discovery solutions

Questions/Comments

dilumb@engr.colostate.edu

www.cnrl.colostate.edu/Projects/CP2P/

CASA is primarily supported by the Engineering Research Centers Program of the National Science Foundation under NSF award number 0313747.